# Training Undergraduates in Genome Analysis and Bioinformatics

Daniel D. Rockey

Department of Biomedical Sciences
College of Veterinary Medicine
Oregon State University
Corvallis, OR

# J. Craig Venter Joint Technology Center

100 machines
~600 Mb / week

Applied Biosystems 3730xl

- 96 capillary array, 2 hr run time
- 50% higher capacity than ABI 3700
- Sequence reads up to 1,100 bases
- Integrated auto-sampler and plate stacker
- Internal bar code reader
- Automated basecalling
- Reduced reagent consumption

- whole genome shotgun sequencing
- PCR resequencing
- full-length + end sequencing of BACs, cDNA libraries

TIGR

Slide: J. Carlton

JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

## Sequencing Progress, Updated Hourly

| Date(s) | Total Q20* Bases | Total Lanes** | % Passed† | Ave. Read Length‡ |
|---|---|---|---|---|
| 9/6/05: ABI3730 | 26.213 Million | 38,592 | 96.05% | 706 |
| 12/16/04: MegaBACE4000 | 3.512 Million | 6,528 | 88.31% | 606 |
| 9/6/05: MegaBACE4500 | 0.221 Million | 384 | 83.59% | 683 |
| Current month (9/05) | 0.358 Billion | 409,632 | 94.93% | 741 |
| Last month (8/05) | 3.012 Billion | 4,182,301 | 94.99% | 756 |
| FY to Date (10/04-9/6/05) | 31.198 Billion | 44,478,145 | 96% | 731 |
| Total (3/99-9/6/05) | 95.201 Billion | 159,385,018 | 91% | 655 |

Real time sequencing statistics from the Joint Genome Institute

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

## Sequencing Progress, Updated Hourly

| Date(s) | Total Q20* Bases | Total Lanes** | % Passed† | Ave. Read Length‡ |
|---|---|---|---|---|
| 9/6/05: ABI3730 | 26.213 Million | 38,592 | 96.05% | 706 |
| 12/16/04: MegaBACE4000 | 3.512 Million | 6,528 | 88.31% | 606 |
| 9/6/05: MegaBACE4500 | 0.221 Million | 384 | 83.59% | 683 |
| Current month (9/05) | 0.358 Billion | 409,632 | 94.93% | 741 |
| Last month (8/05) | 3.012 Billion | 4,182,301 | 94.99% | 756 |
| FY to Date (10/04-9/6/05) | 31.198 Billion | 44,478,145 | 96% | 731 |
| Total (3/99-9/6/05) | 95.201 Billion | 159,385,018 | 91% | 655 |

Real time sequencing statistics from the Joint Genome Institute

JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

## Sequencing Progress, Updated Hourly

| Date(s) | Total Q20* Bases | Total Lanes** | % Passed† | Ave. Read Length‡ |
|---|---|---|---|---|
| 9/6/05: ABI3730 | 28.470 Million | 42,048 | 96.03% | 703 |
| 12/16/04: MegaBACE4000 | 3.512 Million | 6,528 | 88.31% | 606 |
| 9/6/05: MegaBACE4500 | 2.203 Million | 3,456 | 90.86% | 699 |
| Current month (9/05) | 0.363 Billion | 413,088 | 94.92% | 741 |
| Last month (8/05) | 3.012 Billion | 4,182,301 | 94.99% | 756 |
| FY to Date (10/04-9/6/05) | 31.202 Billion | 44,481,601 | 96% | 731 |
| Total (3/99-9/6/05) | 95.204 Billion | 159,391,930 | 91% | 655 |

UWGC

# University of Washington Genome Center

Home | Publications | Technology | Resources | News | Feedback

## HIGH SCHOOL
HUMAN GENOME PROGRAM

HOME | PROGRAM DESCRIPTION | STUDENT ACTIVITIES | TEACHER RESOURCES | PARTICIPATING SITES | RELATED SITES

# Our efforts to train undergraduates

- Two approaches
  - Course in Microbial Genomics
  - Team-based research projects

# Our efforts to train undergraduates

- Two approaches
  - Course in Microbial Genomics
  - Team-based research projects

Microbial Genomics, Summer 2005

2 credits, lecture plus laboratory

Students from Microbiology and Biochemistry

Junior and Senior level.

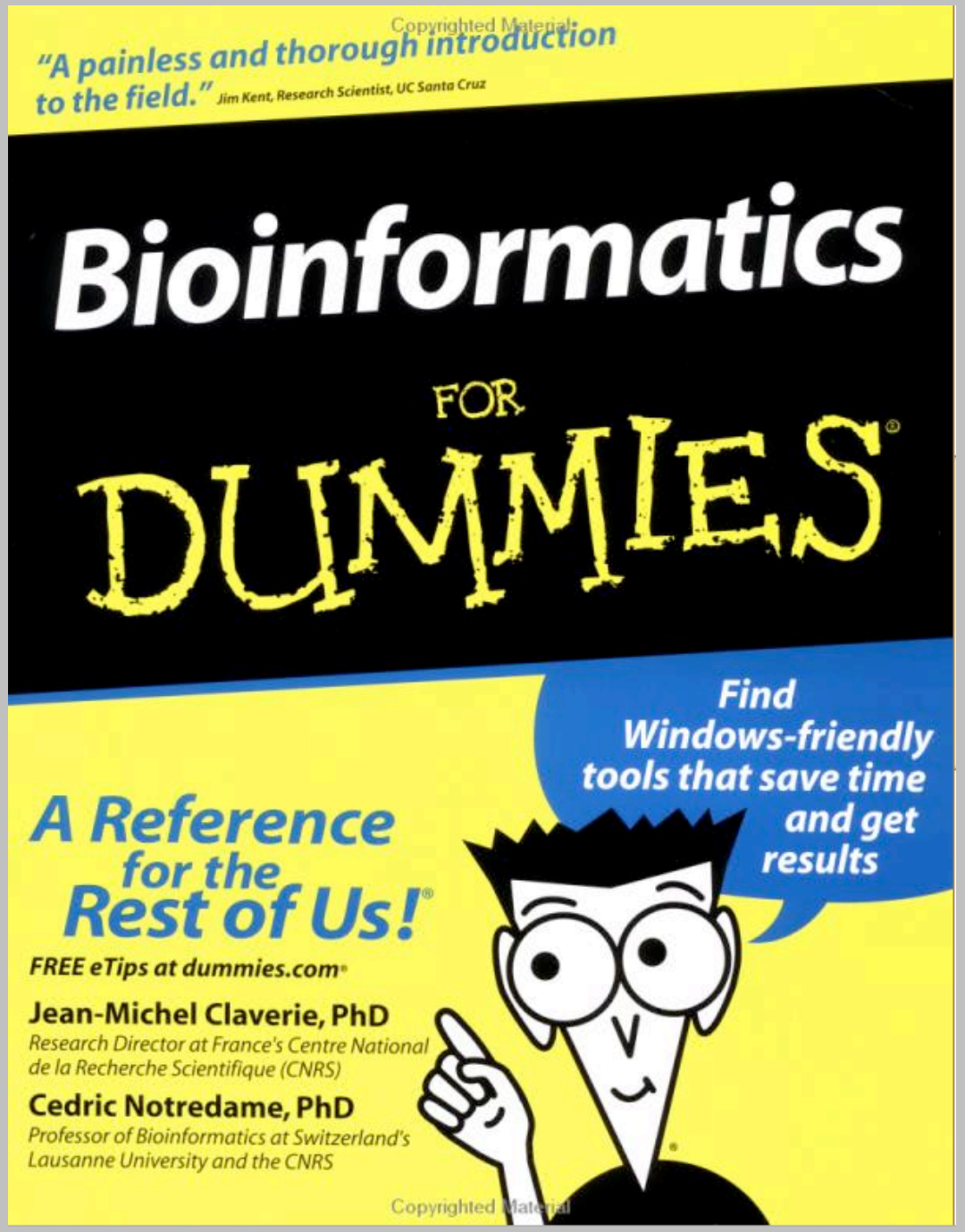
Prereqs: sophomore level biology and a genetics course.

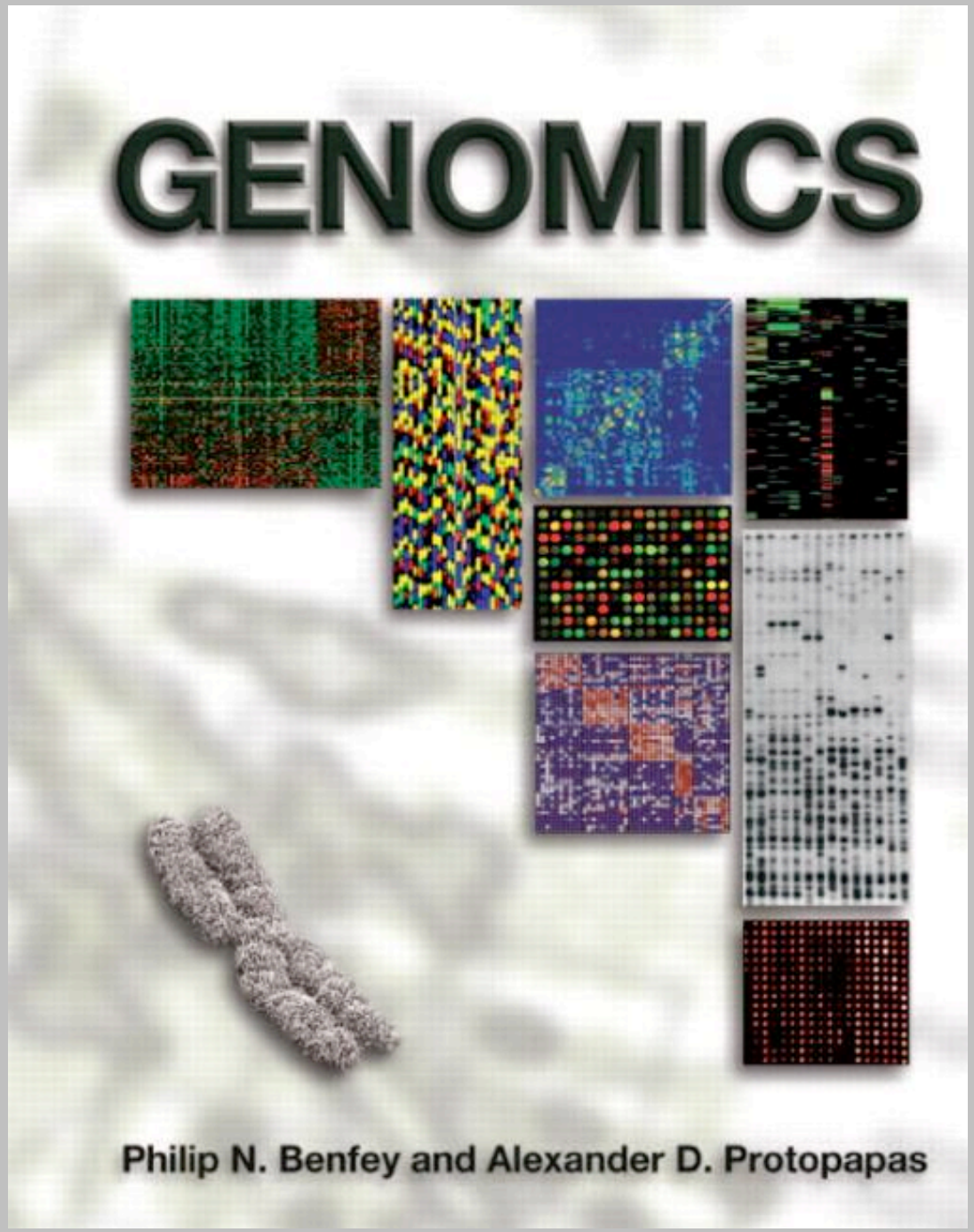
Assessment: Student participation
Student presentation
Take home final exam

# Subject matter covered in the course

Introduction:  what does genomics mean?
How genomes are sequenced.
The computational foundation of genomics.
Gene families and evolution.
Annotation of genomes.
The other “-omics”: genome-based transcription and translation analysis.
Genomic evolution.
Comparative genomics.
Genome structure.
Genomes and medicine.
The future:  systems biology

Choice of textbooks






Powerpoint- based text.

Updates continuously available on the web.

Good overview, lacking in quantitative aspects of bioinformatics.

# Chapter 1

# Introduction

What is genomics, and what are its applications?

## Contents


- Why this format for a genomics "textbook"?
- What is genomics?
- Where did it come from?
- What are its major technologies?
- What are its applications?
- What are the ethical issues?
- Where is it going?


## Why this format for a genomics "textbook"?

- A new format for a rapidly changing field
- Flexibility
  - Slides can be modified and rearranged
- Download updated material
  - URL: www.prenhall.com/benfey
- Modularity
  - Chapters are independent
- Print edition
  - Figures and text
  - Space for notes

## What is genomics?

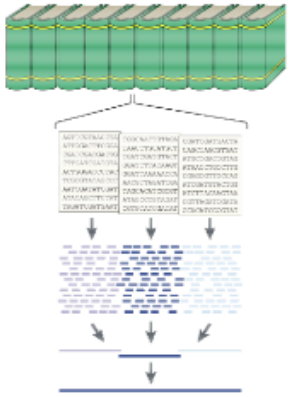
- A marriage of molecular biology, robotics, and computing
- Tools and techniques of recombinant DNA technology
  - e.g., DNA sequencing
- High-throughput technology
  - e.g., robotics for sequencing
- Computers are essential for processing and analyzing the large quantities of data generated

## Genome sequencing

- Analogy: Complete works of an author
  - in partially understood language
- Two approaches
  - Page by page
  - All at once

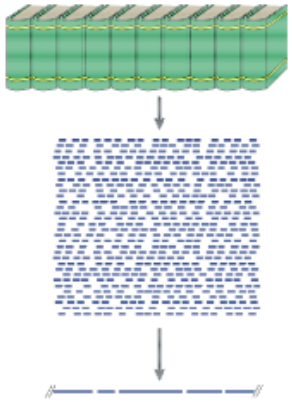
## Page-by-page sequencing strategy

- Sequence = determining the letters of each word on each piece of paper
- Assembly = fitting the words back together in the correct order

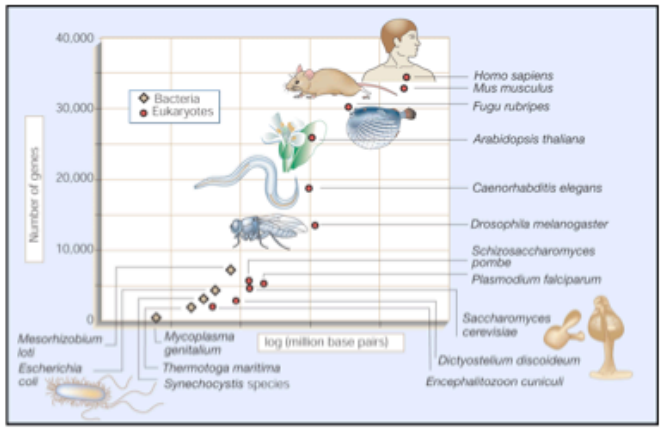


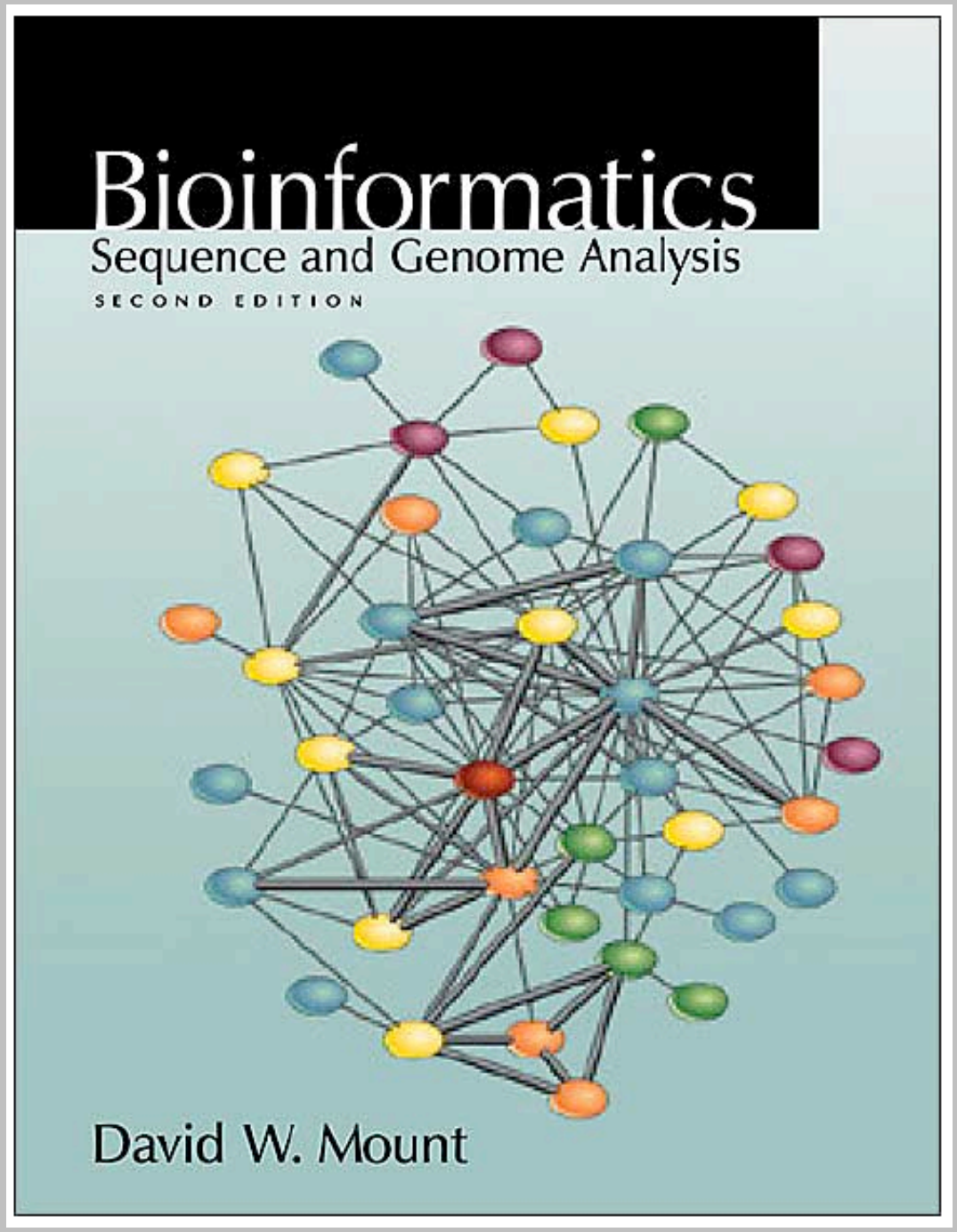
## All-at-once sequencing strategy

- Find small pieces of paper
- Decipher the words on each fragment
- Look for overlaps to assemble



## Genome size and gene number






A second recommended text

Stronger on quantitative aspects

Also update-able on the web.

For bioinformatics texts, Google search:

Listmania! Computational Biology/Bioinformatics

# Student presentations:

- Each student presented a primary research paper outlining the genome sequencing project for a bacterial species.

- The presentations focused on unique aspects of an organism that can be elucidated from the genome
  - Organisms examined this way:
    - Pseudomonas, Wigglesworthia, Agrobacterium, Tropheryma, Nanoarchaeum, Rhodobacter, Staphylococcus, Mycoplasma, Borrelia burgdorferi, E. coli 0157:H7

# Laboratories

- Can’t have too many laboratories- computer based analysis of sequence.
- Conduct basic BLAST analysis
- Examine annotation, phylogeny strategies

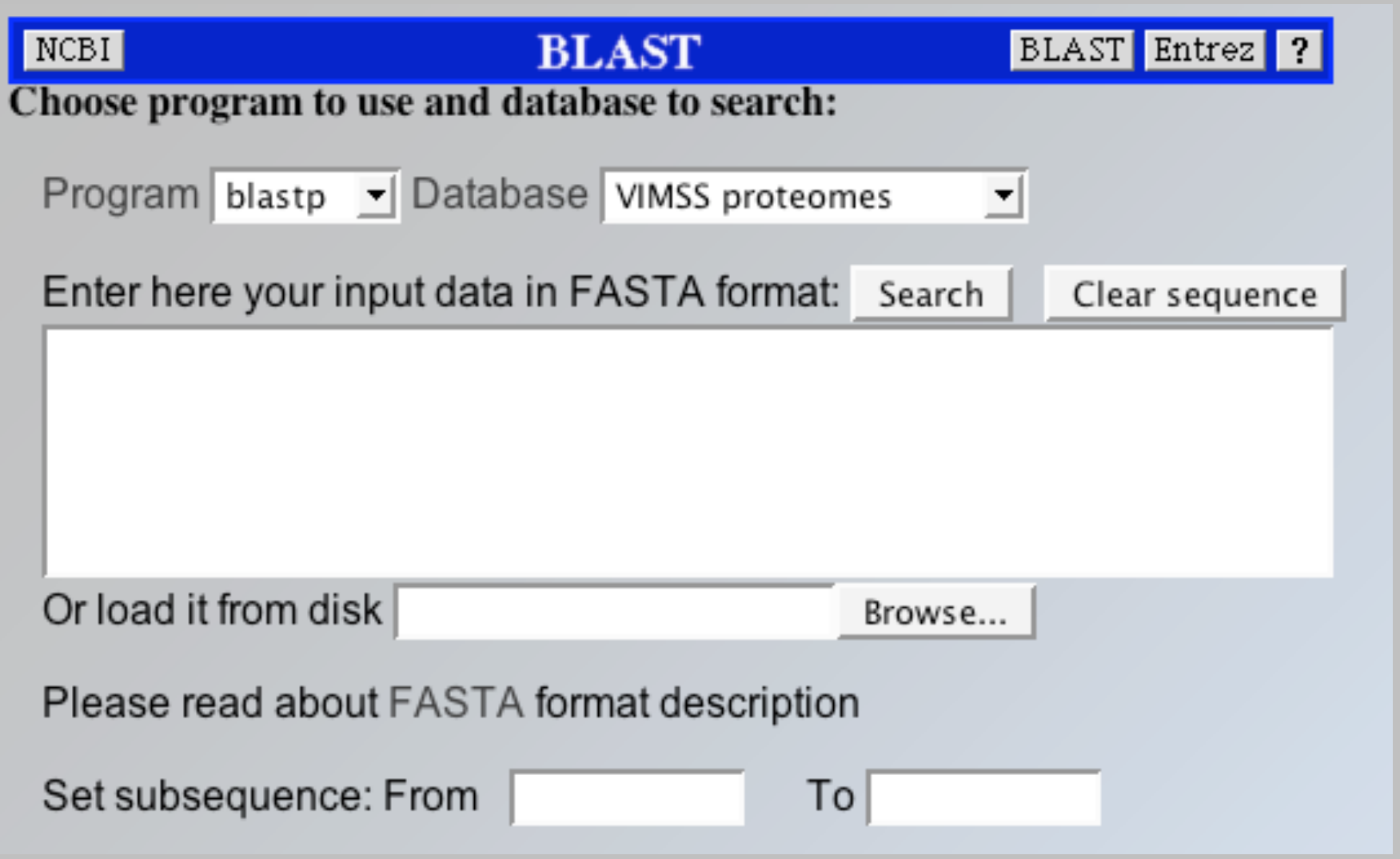
- Goal of the Laboratory: Practical training in genomics analysis.

- Computer based laboratories based on the genomes present in the Ergo Light program, as established by Integrated Genomics (http://www.ergo-light.com/ERGO/).

- We attempted to base our analyses on the genome sequence of *Renibacterium salmoninarum*- became challenging as all the data weren't in or finalized.

- Used some data from the Sar11 genome- based on practical challenges faced by Jim Tripp as he annotated the genome.

# Introductory lab: BLAST analysis



# Example of laboratory

## Gene Finding and Comparative Genomics Lab

Lab written by Jim Tripp, graduate student, Oregon State University

My laboratory needs your help. Having received a $3.23 million dollar grant from the Moore Foundation (billionaire founder of Intel and inventor of "Moore's Law") our group Lab provided several strains for genomic sequencing and annotation. The sequencing is being done on an additional two-year $8.95 million dollar grant by the J. Craig Venter Institute, (inventor of genome "shotgunning") which is highly acclaimed for sequencing ability and not so well acclaimed for genome annotation and comparison. Our laboratory has six months to review the sequences, make comparisons, write Science and Nature papers, and collaborate with the Venter Institute on final annotations to be released to the public and scientific community.

**Welcome**

Home
Status

**Organisms**

*C. atlanticus*
*E. litoralis*
*F. bacterium*
*F. pelagi*
*γ-proteobacteria*
*Janibacter sp.*
*L. araneosa*
*O. alexandrii*
*O. batsensis*
*O. granulosus*
*P. bermudensis*
*P. ubique 1002*
*P. ubique 1062*
*P. ubique 1062hc*
*R. bacterium*
*R. biformata*

# Organism Status

| Strain ID | Genus Species | Annotations | KEGG | Syntigs | GBrowse | COGs |
|---|---|---|---|---|---|---|
| HTCC2559 | *Croceibacter atlanticus* | Yes | Yes | Yes | Yes | Yes |
| HTCC2594 | *Erythrobacter litoralis* | Yes | Yes | - | Yes | Yes |
| HTCC2170 | *Flavobacteriales bacterium* | Yes | Yes | - | Yes | Yes |
| HTCC2506 | *Fulvimarina pelagi* | - | - | - | - | - |
| HTCC2207 | *Gammaproteobacteria sp.* | Yes | Yes | - | Yes | Yes |
| HTCC2649 | *Janibacter sp.* | Yes | Yes | Yes | Yes | Yes |
| HTCC2155 | *Lentisphaera araneosa* | - | - | - | - | - |
| HTCC2633 | *Oceanicaulis alexandrii* | Yes | Yes | Yes | Yes | Yes |
| HTCC2597 | *Oceanicola batsensis* | Yes | Yes | Yes | Yes | Yes |
| HTCC2516 | *Oceanicola granulosus* | Yes | Yes | - | Yes | Yes |
| HTCC2503 | *Parvularcula bermudensis* | Yes | Yes | Yes | Yes | Yes |
| HTCC1002 | *Pelagibacter ubique* | Yes | Yes | - | Yes | Yes |
| HTCC1062 | *Pelagibacter ubique* | Yes | Yes | - | Yes | Yes |
| HTCC1062 | *Pelagibacter ubique (hand curated)* | Yes | Yes | Yes | Yes | Yes |
| HTCC2654 | *Rhodobacterales bacterium* | Yes | Yes | - | Yes | Yes |
| HTCC2501 | *Robiginitalea biformata* | Yes | Yes | - | Yes | Yes |

Gene synteny is very useful for filling in holes in operons. A great tool for examining synteny is Microbes On-Line. The information in the "About MicrobesOnline" box is very useful, especially their Comparative Genome Browser, which can be linked to after getting the results from their particular BLAST tool.

Holes in metabolic pathways are best studied using the CGRB Adaptation of KEGG. Once you get a display of the pathway in question, (you will study histidine biosynthesis), clicking on present and absent enzymes will give you examples of that enzyme from other organisms if your organism doesn't have it, or it will take you to a display of the ORF in your organism that does have it.

Problem: Identify the missing genes in the apparently incomplete histidine biosynthesis pathway in *Oceanicaulis alexandrii*

http://gac-web.cgrb.oregonstate.edu:8080/microbes/kegg/viewPathway.html

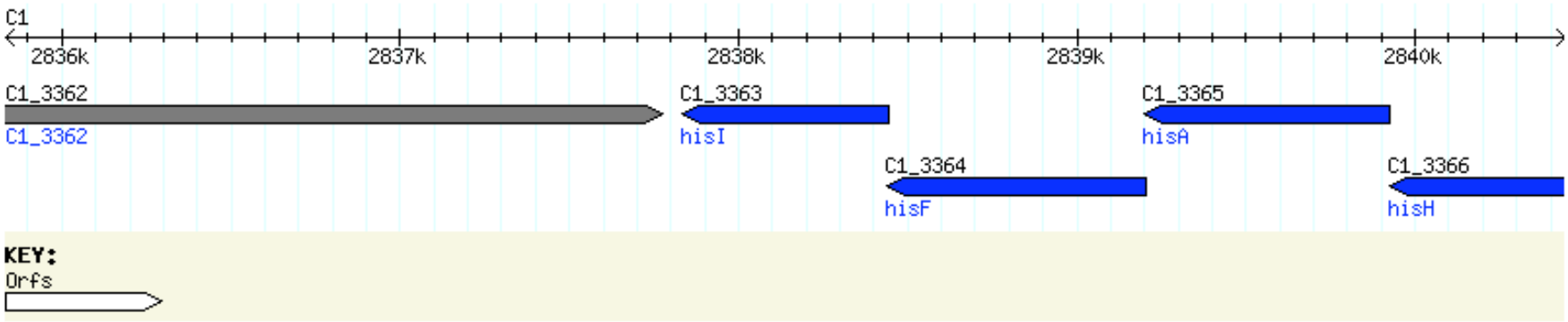
# Annotation Search Engine

## Organism: *Oceanicaulis alexandrii*

| Search Term | Search Type | Organism | |
|---|---|---|---|
| | BLAST Hits | Oceanicaulis alexandrii | Submit |

## Gene: hisI

### ORF: C1_3363




Show All Facts  Show All Annotations

### Annotation

| | |
|---|---|
| Product: | (Q87C35) Histidine biosynthesis bifunctional protein hisIE [Includes: Phosphoribosyl-AMP cyclohydrol |
| Description: | (Q87C35) Histidine biosynthesis bifunctional protein hisIE [Includes: Phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) (PRA-CH); Phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) (PRA-PH)] |
| Comment: | auto-annotation derived from facts Tool: BLASTP of Swissprot + PIR DB ID: HIS2_XYLFT |
| EC: | 3.5.4.19 |

Problem: Missing ribosomal proteins. The primary annotation of Croceibacter atlanticus indicated that the organism lacks the following gene products.

- COG0227 Ribosomal protein L28
- COG0255 Ribosomal protein L29
- COG0048 Ribosomal protein S12
- COG0267 Ribosomal protein L33
- COG0093 Ribosomal protein L14
- COG0333 Ribosomal protein L32

Are these really missing, or did the pipeline simply fail to detect them?

# Our efforts to train undergraduates

- Two approaches
  - Course in Microbial Genomics
  - Team-based research projects

# Team-based research

- Teams of students, some at OSU, some at UW.
- Examine technically repetitive aspects of the genome
- Cross check results both within teams and between teams.

# Bioinformatic exploration of *Renibacterium salmoninarum*


**Principal Investigators:**

**Dr. Dan Rockey**

**Dr. Mark Strom**

**Dr Greg Wiens**

**Undergraduate investigators:**

**Windy Beck**

**Nathaniel Blake**

**Gina Capri**

**Bob Ruef**

Oregon State University OSU


•Project funded by USDA

ERGO™

BIOINFORMATICS SUITE
comparative genomics analysis
by Integrated Genomics, Inc.

Integrated Genomics



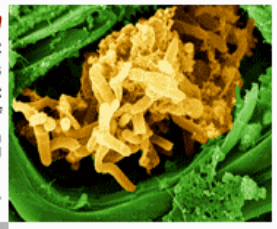
Wednesday April 6, 2005

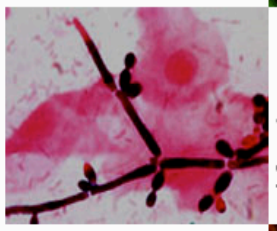
**ERGO™ bioinformatics suite by Integrated Genomics, Inc.**
is an integration of biological data
from genomics, biochemistry, genetics
and high-throughput expression profiling,
to achieve a comprehensive analysis of genes and genomes

### *Xylella fastidiosa*

*Xylella fastidiosa* causes wilt disease in plants and is responsible for major economic and crop losses. The genome of *Xylella fastidiosa* has been analyzed using ERGO™

Proc Natl Acad Sci USA 2002;99(19):124





### *Candida albicans*

*Candida albicans* is one of the most commonly encountered human pathogens, causing a wide variety of infections

Genome Source: Stanford Genome Technology Center

### *Glycine max*

One of the world's most important oilseed crops and a staple food of the Orient. Indigenous to the Far East, it is now cultivated elsewhere

Genome Source: IG & University of Minnesota

**Registered user, please enter your user name and password:**

User name

Password

Enter ERGO

**To get an access to ERGO™, please visit our "Subscription" section**

## News

**Dec 13, 2004** New genomes added into ERGO: *Burkholderia mallei ATCC 23344*, *Chlorochromatium aggregatum*, *SARS coronavirus*

**Oct 15, 2004** Patricia Langer, Anamitra Bhattacharyya, Vinayak Kapatral, Kathe Andrews-Cramer 'High-throughput Analysis of Microbial Genome Data' (in German), Biospektrum 2004, No. 5: 691-692

## Statistics

Number of complete (■) and gapped (■) genomes, available through ERGO™:

| | |
|---|---|
| Bacteria | 281/114 |
| Archaea | 24/8 |
| Viral | 169/3 |
| Eukarya | 25/92 |

# Protein Page for RRSA01105 from Renibacterium salmoninarum (IG-152)

Save Page

**Data Panel Display**

Select Data Panel

**Examine RRSA01105**

- View Annotations
- Local Blast (NR) -- Protein
- Local Blast (NR) -- DNA
- Function Cluster
- Function Couplings
- Function Tree
- Possible Fusion Event
- Paralogs
- Paralog Cluster

**External Tools**

- TMPred
- REBASE
- ProtScale
- PSI-Blast (NR)
- RPS-Blast (NR)
- EBI NCBI BLAST2
- ProSite
- ProDom Analysis

## Primary Information for RRSA01105 Renibacterium salmoninarum (IG-152)

| | |
|---|---|
| Aliases | None |
| Contig Location | Contig0210 from 6,472 to 6,023; contig length = 18,574 bp |
| AA Residues, DNA | 150 aa, 450 bp |
| Amino acid count | A 18, C 1, D 4, E 3, F 13, G 20, H 3, I 9, K 8, L 16, M 3, N 4, P 5, Q 4, R 5, S 4, T 9, V 15, W 1, Y 5 |
| Molecular Weight | 15,944 Da |
| Iso-electric Point | 10.14 |
| GC content | 54.67%, entire genome value = 56.24%, difference = -1.57% |
| Function | Beginning of transporter, MFS superfamily |
| Master Function | |

## Contig Region for RRSA01105

Neighboring Genes

## Pathway Information for RRSA01105

No Pathways Associated with *Beginning of transporter, MFS superfamily*

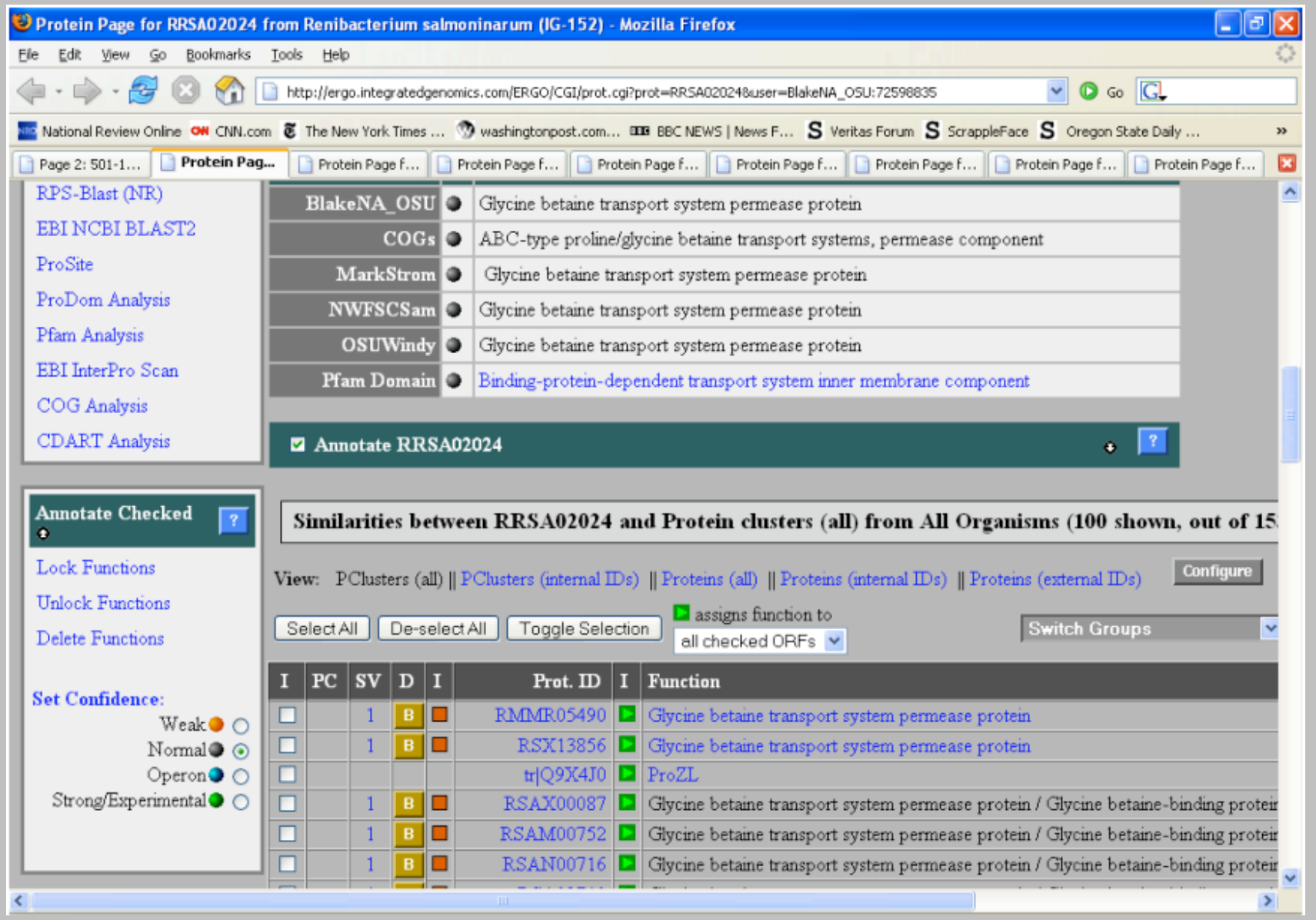
## External Annotations for RRSA01105

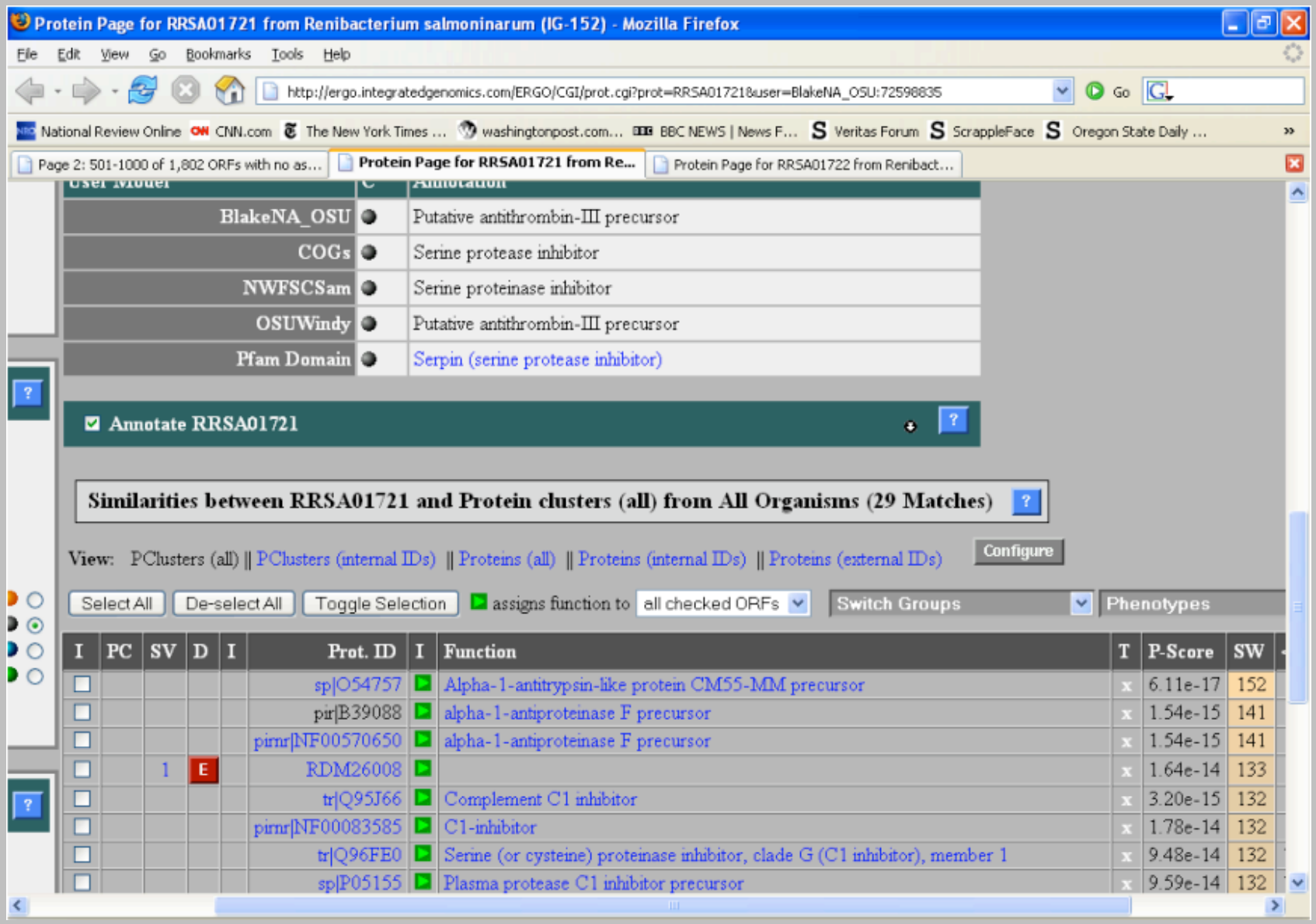
| User Model | C | Annotation |
|---|---|---|
| BlakeNA_OSU | | beginning of transporter, MFS superfamily |
| COGs | | Permeases of the major facilitator superfamily |

# An Amino Acid Alignment Between Target Organism and a Selected Gene for Similarity Assessment at the Peptide Level

This Amino Acid alignment below is generated with ClustalW, an excellent alignment program written by Desmond Higgins and coworkers. We are most grateful for being able to use it.

```
CLUSTAL W (1.7) multiple sequence alignment


RZC00359        MINKTLLQKRFNGAAVSYDRYANVQKKMAHSLLSILKERYSETASIRILELGCGTGYVTE
RBAT05226       MINKTLLQKRFNVAAVSYDQYANVQKKMAHSLLSTLDRRYSANSSIRILELGCGTGYVTE
                *********** ******:************** *..*** .:*****************

RZC00359        QLSKLFPKSHITAVDFAESMIAIAQTRQNVKNVTFHCEDIERLRLEESYDVIISNATFQW
RBAT05226       QLSNLFPKAHITAIDFAESMIAVAKTRQNVKNVMFYCEDIERLQLEETYDVIISNATFQW
                ***:****:****:********:*:******** *:*******:***:************

RZC00359        LNNLQQVLRNLFQHLSIDGILLFSTFGHETFQELHASFQRAKEERNIKNETSIGQRFYSK
RBAT05226       LNDLKQVIRNLFHHLSIDGILLFSTFGQETFQELHTSFQRAKEEKNIQNETSIGQRFYSK
                **:*:**:****:**************:*******:********:**:************

RZC00359        DQLLHICKIETGDVHVSETCYIESFTEVKEFLHSIRKVGATNSNEGSYCQSPSLFRAMLR
RBAT05226       NQLRHICEVETGDVHVSETCYIERFTEVREFLHSIRKVGATNSNEESYCQSPSLFRAMLR
                :** ***::************** ****:***************** **************

RZC00359        IYERDFTGNEGIMATYHALFIHITKEGKR
RBAT05226       IYERDFTGNEGIMATYHALFMHITKEGKR
                *********************:********
```


Protein Page for RRSA02024 from Renibacterium salmoninarum (IG-152) - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://ergo.integratedgenomics.com/ERGO/CGI/prot.cgi?prot=RRSA02024&user=BlakeNA_OSU:72598835
Go
National Review Online CNN.com The New York Times ... washingtonpost.com... BBC NEWS | News F... Veritas Forum ScrappleFace Oregon State Daily ...
Page 2: 501-1... Protein Pag... Protein Page f... Protein Page f... Protein Page f... Protein Page f... Protein Page f... Protein Page f... Protein Page f...
RPS-Blast (NR)
EBI NCBI BLAST2
ProSite
ProDom Analysis
Pfam Analysis
EBI InterPro Scan
COG Analysis
CDART Analysis
BlakeNA_OSU Glycine betaine transport system permease protein
COGs ABC-type proline/glycine betaine transport systems, permease component
MarkStrom Glycine betaine transport system permease protein
NWFSCSam Glycine betaine transport system permease protein
OSUWindy Glycine betaine transport system permease protein
Pfam Domain Binding-protein-dependent transport system inner membrane component
Annotate RRSA02024
Annotate Checked
Lock Functions
Unlock Functions
Delete Functions
Set Confidence:
Weak
Normal
Operon
Strong/Experimental
Similarities between RRSA02024 and Protein clusters (all) from All Organisms (100 shown, out of 15
View: PClusters (all) || PClusters (internal IDs) || Proteins (all) || Proteins (internal IDs) || Proteins (external IDs)
Configure
Select All De-select All Toggle Selection
assigns function to all checked ORFs
Switch Groups
I PC SV D I Prot. ID I Function
RMMR05490 Glycine betaine transport system permease protein
RSX13856 Glycine betaine transport system permease protein
tr|Q9X4J0 ProZL
RSAX00087 Glycine betaine transport system permease protein / Glycine betaine-binding protein
RSAM00752 Glycine betaine transport system permease protein / Glycine betaine-binding protein
RSAN00716 Glycine betaine transport system permease protein / Glycine betaine-binding protein


Protein Page for RRSA01721 from Renibacterium salmoninarum (IG-152) - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://ergo.integratedgenomics.com/ERGO/CGI/prot.cgi?prot=RRSA01721&user=BlakeNA_OSU:72598835
Go
National Review Online CNN.com The New York Times ... washingtonpost.com... BBC NEWS | News F... Veritas Forum ScrappleFace Oregon State Daily ...
Page 2: 501-1000 of 1,802 ORFs with no as...
Protein Page for RRSA01721 from Re...
Protein Page for RRSA01722 from Renibact...
BlakeNA_OSU Putative antithrombin-III precursor
COGs Serine protease inhibitor
NWFSCSam Serine proteinase inhibitor
OSUWindy Putative antithrombin-III precursor
Pfam Domain Serpin (serine protease inhibitor)
Annotate RRSA01721
Similarities between RRSA01721 and Protein clusters (all) from All Organisms (29 Matches)
View: PClusters (all) || PClusters (internal IDs) || Proteins (all) || Proteins (internal IDs) || Proteins (external IDs)
Configure
Select All
De-select All
Toggle Selection
assigns function to all checked ORFs
Switch Groups
Phenotypes
I PC SV D I Prot. ID I Function T P-Score SW
sp|O54757 Alpha-1-antitrypsin-like protein CM55-MM precursor x 6.11e-17 152
pir|B39088 alpha-1-antiproteinase F precursor x 1.54e-15 141
pirnr|NF00570650 alpha-1-antiproteinase F precursor x 1.54e-15 141
1 E RDM26008 x 1.64e-14 133
tr|Q95J66 Complement C1 inhibitor x 3.20e-15 132
pirnr|NF00083585 C1-inhibitor x 1.78e-14 132
tr|Q96FE0 Serine (or cysteine) proteinase inhibitor, clade G (C1 inhibitor), member 1 x 9.48e-14 132
sp|P05155 Plasma protease C1 inhibitor precursor x 9.59e-14 132

# Assessment- how successful were we?

- Two OSU students have gone to different research appointments
- OSU undergrads presented material at a conference
- Active assessment of course objectives is in development
- What we will add:  more actual bench time
  - Team scientists did southern blots, students  in the genomics class did no bench work

Acknowledgements

Greg, Mark

Jim Tripp, OSU

Windy Beck
Nathanael Blake
Rob Ruef
Gina Capri

Sam Crane, Donald Chen

USDA/CSREES/NSF support for the project